学校编号： 10384　　分类号：______密级：______
学　号： 200323033　　UDC:______

硕 士 学 位 论 文

广义指数族 ARMA 模型的参数估计及其 BOOTSTRAP 置信区间

Parameters Estimate and Bootstrap Confidence Intervals of Generalized Exponential Family ARMA Models

陈 �londa 筠

指导教师姓名： 林建华　　教授

厦门大学数学科学学院

申请学位级别： 硕　士　学　位

专 业 名 称： 概率论与数理统计

论文提交日期： 2006 年 4 月

论文答辩日期： 2006 年 6 月

学位授予单位： 厦　门　大　学

学位授予日期： 2005 年 6 月

答辩委员会主席：______

评阅人：______

2006 年 4 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文而产生的责任。

声明人 (签名):

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（ ）在　　年解密后适用本授权书。

2、不保密。

（请在以上相应括号内打"√"）

作者签名：　　　　　　　　　年　　月　　日

导师签名：　　　　　　　　　年　　月　　日

目　录

Abstract

摘　要

本文针对广义指数族 ARMA 模型，分为连续型指数族和离散型指数族两种情况讨论模型的参数估计，并运用不同的 Bootstrap 方法构造参数的置信区间和置信带. 两种情况均采用 Scoring 算法进行模型的参数估计，并分别得到 Scoring 算法中两个模型方向向量的计算公式. 对于连续型情况，运用 Wild Bootstrap 构造参数的置信区间和置信带，并将其结果与传统的残差 Bootstrap 进行比较，得到 Wild Bootstrap 更快更精确的结论；对于离散型情况，运用分块移动 Bootstrap 构造参数的置信区间，这种方法更加实用，收敛速度快，并得到较为满意的结果.

关键词：广义指数族 ARMA 模型； Scoring 算法； Wild Bootstrap ；分块移动 Bootstrap.

Abstract

The objective of this paper is to estimate the parameters of generalized exponential family ARMA models and construct the confidence intervals of the parameters. In this paper we divided the problem into continuous type situation and discrete type situation. Under both situations, we use Fisher Scoring algorithm to estimate the parameters and gain the formula of direction vectors in the algorithm. Under the continuous situation we use Wild Bootstrap to construct the confidence intervals of the parameters,then compare the simulation results to the Remain Bootstrap, then get the conclusion that the Wild Bootstrap is more quickly and more accurate; Under the discrete situation the method of Moving Blocks Bootstrap gives satisfied results for both simulative and true data.

Keywords: Generalized exponential family ARMA models; Scoring Algorithm; Wild Bootstrap; Moving Blocks Bootstrap.

第一章 绪 论

§1.1 背景知识

§1.1.1 ACD 模型的背景

近年来, 计算工具与计算方法的发展, 极大地降低了数据记录和存储的成本, 使得对大规模数据库的分析成为可能. 所以，许多科学领域的数据都开始以越来越精细的时间刻度来收集，这样的数据称为高频数据 (high frequency data). 金融市场中, 逐笔交易数据 (transaction by transaction data) 或逐秒记录数据 (tick by tick data) 就是高频数据的例子.

金融高频数据的一类重要分支是关于金融高频数据分析中所使用的计量模型的研究. 随着金融高频数据的不断增加，如何使用模型来恰当地描述这些数据就成为一个重要的问题. 从计量经济学角度来看，金融高频数据的一个最显著特征是观测值以变动的、随机的时间间隔取得，即金融高频数据是一个动态的时间序列. 在过去的实证研究中，我们采用的是固定时间间隔的计量模型 (如关于波动性研究的 GARCH 模型和 SV 模型等) 的方法，忽略了交易数据之间存在的固有的时间间隔. 然而，就象金融市场微观结构理论模型所描述的那样，不同交易之间的时间间隔蕴含着大量交易者行为的信息. 因此，如果采用固定时间间隔的计量模型方法，实证结果将有可能丢失大量的高频交易中所包含的信息成分.

Engle&Russell(1998)[1] 针对这个问题使用了与分析波动性的 ARCH 模型相似的概念提出了 ACD(Autoregressive Conditional Duration) 模型来描述 (交易活跃的) 股票交易间隔的发展过程. Engle&Russell 把交易之间的时间间隔当作随机变量并建立计量模型，这些随机变量遵循一个点过程 (point process). 交易数据可以用两个随机变量来描述：第一个是交易发生的时间；第二个是在交易发生时间上所观察到的一个向量，这个向量被称为记号 (Marks) ，用来确定所研究

的事件 (Events) 是否发生了. 通俗地说, 就是在时间这个尺度上用所研究的事件当作刻度来进行标记. 这个标记可以是交易量、成交价格或买卖价差等反映投资者交易情况的事件，而这些正是目前金融市场微观结构理论研究的重点.

Engle 在 ACD 模型的研究中所起作用等同于他对 ARCH 模型发展的贡献. 更确切地说， Engle(2000)[2] 可以看作是对高频数据计量分析的宣言书. ACD 模型完善于 Engle&Rusell (1998)[1] 在 Econometrica 上发表的文章. 他们的思想是在原有的 ARCH 模型的框架下, 用一个标记点过程 (marked point process) 去刻画随机交易的时间间隔，不同的点过程假设自然就得到了不同的 ACD 模型. Engle&Rusell(1996)[3] 用 ACD 模型成功地完成对交易频率等实时交易变量的预测并将类似门限的思想引入后，提出了一种非线形的 ACD 模型. 他们的模型后来在 Zhang et al(1999)[4] 中进一步得到了拓展.

§1.1.2 ACD 模型的定义

ACD 模型是在过去事件的基础上为分析研究交易持续期 (duration) 的条件分布而创建的，这个模型把交易的持续期（时间间隔）转化为一个随时间间隔变动的动态的点过程.

为了进一步说明这个问题, 我们考虑交易的抵达时间序列 $\{t_0, t_1, \cdots, t_n, \cdots\}$ 其中 $t_0 < t_1 < \cdots < t_n < \cdots$，则交易持续期定义为

$$\Delta t_i = t_i - t_{i-1}$$

它是两个事件间发生的时间间隔. 我们知道在证券市场上一天中的交易有时密集有时稀少，这样交易之间的持续期是动态变动的. 为了得到一个独立的变量，必须对一天中的交易持续期进行调整 (An Interday Seasonal Adjustment)，剔除一天中交易时间段对交易持续期的影响. 比如，在开盘时交易比较密集，而在进入连续竞价交易后交易有可能比较稀少. 因此，为了除去这种交易时间效应对交

易持续期的影响，进行了如下调整：令

$$x_i = \Delta t_i^* = \frac{\Delta t_i}{f(t_i)}$$

为调整后的交易持续期，其中 $f(t_i)$ 是 Δt_i 的周期性成分构成的一个确定性 (deterministic) 函数，通常取 $f(t_i)$ 为平滑样条 (smoothing spline).

对于调整后的交易持续期 x_i , 定义 ψ_i 为第 i 个交易持续期 x_i 在 $x_1, \cdots, x_{i-1}$ 下的条件期望值：

$$\psi_i \equiv \psi_i(x_{i-1}, \cdots, x_1; \theta) = E(x_i | x_{i-1}, \cdots, x_1) \tag{1.1}$$

其中 θ 为自由变量，并假设：

$$x_i = \psi_i \cdot \varepsilon_i \tag{1.2}$$

其中

$$\varepsilon_i \sim i.i.d.\ \Re(\phi) \tag{1.3}$$

为模型的误差项， $\Re$ 为带有自由变量 ϕ 的某个分布，且 $E(\varepsilon_i) = 1$.

由于持续期的条件期望与过去的持续期有关，称此模型为自回归条件持续期 (Autoregressive Conditional Duration) 模型，简称为 ACD 模型. 由公式 (1.1)-(1.3) 易见根据交易持续期的条件期望 ψ_i 的表达形式和 ε_i 分布的不同， ACD 模型有大量不同的表达形式.

ACD 模型中一个简单形式是条件期望值 ψ_i 仅以过去的持续期作为条件，即第 i 个持续期的条件期望是由其滞后 p 个过去的持续期和滞后 q 个的条件期望共同来决定，可以表述为如下方程：

$$\psi_i = \alpha_0 + \sum_{m=1}^{p} \alpha_m x_{i-m} + \sum_{n=1}^{q} \beta_n \psi_{i-n} \tag{1.4}$$

称为 ACD(p,q) 型，其中 p 和 q 表示滞后的阶数.

这类模型的优点在于易于计算出模型的各阶矩, 例如, x_i 的条件期望为 ψ_i , 而 x_i 的无条件期望 $\overline{\mu}$ 为

$$\overline{\mu}=E(x_i)=\frac{\omega}{1-\sum\limits_{j=1}^{max(p,q)}(\alpha_j+\beta_j)}$$

其中当 $j>p$ 时, $\alpha_j=0$; 而当 $j>q$ 时, $\beta_j=0$, 上式只需对 (1.4) 式两边同时取期望即可得到.

这类模型中最简单的是 EACD(1,1) 模型, 其中 E 表示式 (1.3) 中的误差项 $\{\varepsilon_i\}$ 为指数分布:

$$\psi_i=\alpha_0+\alpha_1x_{i-1}+\beta_1\psi_{i-1}$$

其中 $\alpha_1,\beta_1\geq 0,\alpha_0>0,\forall i$. 这个模型中 x_i 的条件方差是 ψ_i^2, 而无条件方差为

$$\sigma^2=\overline{\mu}^2\left(\frac{1-\beta_1^2-2\alpha_1\beta_1}{1-\beta_1^2-2\alpha_1\beta_1-2\alpha_1^2}\right)$$

与 GARCH 模型相似, 模型 (1.4) 也有 ARMA 表达形式. 令 $\eta_i\equiv x_i-\psi_i$ 为一个鞅差序列, 则持续期过程可以表示为:

$$x_i=\alpha_0+\sum_{m=1}^{max(p,q)}(\alpha_m+\beta_m)x_{i-m}-\sum_{n=1}^{q}\beta_n\eta_{i-n}+\eta_i$$

上式为具有非高斯创新的 ARMA 表达形式. 因此模型的预测问题可以通过 ARMA 方法完成.

§1.1.3 ACD 模型的扩展

过去几年中, 对 ACD(p,q) 模型的研究取得了大量的成果. 其中最普遍的是假设公式 (1.3) 中的 $\Re$ 为 Weibull 分布, 另一趋势是假设 $\Re$ 为广义 Gamma 、对数 Logistic 以及对数正态分布. 由于指数族包含许多重要的分布, 如正态分布、伽玛分布、泊松分布、二项分布等, 本文针对指数族的情形进行讨论.

当 x_i 为连续型随机变量时，采用原有的 ACD 模型，并假设误差项 $\{\varepsilon_i\}$ 的分布 $\Re$ 为连续型指数族分布，本文称此种情况下的模型为广义连续型指数族 ARMA 模型.

然而现实中由于消费者的持续购买力各不相同, 上述持续期 x_i 可能是数月甚至数年, 从而在实证分析中该类数据的取样会出现离散的情况. 故本文就 x_i 为离散型随机变量的情况将模型改进如下:

假设一维随机变量 x_i ，令 $F_{i-1} = \sigma(x_{i-1},\cdots,x_1,\psi_{i-1},\cdots,\psi_1)$ 为过去第 $i-1$ 个交易过程提供的信息生成的 σ 代数，其中 ψ_i 为 x_i 的条件期望，条件密度 $f(x_i|F_{i-1})$ 为非负离散型指数族，条件期望 $\psi_i = E(x_i|F_{i-1})$ 的形式为

$$\psi_i = h\left(\alpha_0 + \sum_{m=1}^{p} \alpha_m x_{i-m}\right) + \sum_{n=1}^{q} \beta_n \psi_{i-n} \tag{1.5}$$

其中 $h(\cdot)$ 为关联 (link) 函数. 本文称上述模型为广义离散型指数族 ARMA(p,q) 模型.

§1.2 文章结构

本文分别对连续型和离散型两种情况讨论模型的参数估计, 运用不同的 Bootstrap 方法构造参数的置信区间和置信带. 本文结构如下: 第二章介绍模型的参数估计方法 (Fisher Scoring 算法) 和 Bootstrap 分位数置信区间的构造方法; 第三章讨论广义连续型指数族 ARMA 模型的参数估计，得到 Scoring 算法中该模型方向向量的计算公式，并运用 Wild Bootstrap 方法构造置信区间和置信带，数据模拟部分将其结果与传统的残差 Bootstrap 做比较，得到 Wild Bootstrap 方法更快更精确的结论；第四章讨论广义离散型指数族 ARMA 模型的参数估计，亦得到 Scoring 算法中该模型方向向量的计算公式，并运用分块移动 bootstrap 方法构造置信区间，然后进行数据模拟和真实数据分析；最后第五章为结论.

第二章 模型的参数估计及其置信区间的构造

§2.1 Fisher Scoring 算法

本文采用极大似然估计法和 Fisher Scoring 算法进行模型的参数估计. Fisher Scoring 算法与 Newton- Raphson 算法类似，其区别在于 Fisher Scoring 算法采用的是模型的二阶微分矩阵的期望值. 其具体步骤如下：

Fisher Scoring 算法

第一步：对待估参数向量赋予初值 $\omega^{(0)}$ ；

第二步：进行 $\widehat{\omega}^{(i)}$ 的迭代计算，参数 ω 的第 $i+1$ 步迭代的估计值为

$$\begin{aligned}\widehat{\omega}^{(i+1)} &= \widehat{\omega}^{(i)}+\lambda_i\cdot I(\widehat{\omega}^{(i)})^{-1}\cdot g(\widehat{\omega}^{(i)})\\ &= \widehat{\omega}^{(i)}+\lambda_i\cdot\left[-E\left(\sum_{i=1}^{N}\frac{\partial^2 l_i}{\partial\omega\partial\omega'}\right)\right]^{-1}\cdot\sum_{i=1}^{N}\frac{\partial l_i}{\partial\omega}|_{\omega=\widehat{\omega}^{(i)}}\end{aligned}$$

其中 l_i 为对数似然函数， $\left[-E\left(\sum_{i=1}^{N}\frac{\partial^2 l_i}{\partial\omega\partial\omega'}\right)\right]^{-1}\cdot\sum_{i=1}^{N}\frac{\partial l_i}{\partial\omega}|_{\omega=\widehat{\omega}^{(i)}}$ 为第 i 步方向向量， $I(\widehat{\omega})$ 为信息矩阵， $g(\widehat{\omega})$ 为 score 向量， λ_i 第 i 步的搜索步长，一般取使在该搜索方向上对数似然函数极大化的步长；

第三步：重复第二步，直到所得参数的估计结果满足一定条件，如收敛性、迭代次数等.

§2.2 Bootstrap 方法

§2.2.1 Bootstrap 方法的介绍

Bootstrap 是近代才发展起来的用于处理某类统计推断的新方法. 这是由于它必须依赖现代计算机技术对传统理论中过于复杂的计算作为支撑. Bootstrap 方法和其他一些以计算机为基础的方法一样其基本思想没有改变，只是执行计算的方法改变了. Bootstrap 已经广泛应用于模型选择、平滑参数数据适应性选择

(如带宽选择) 、检验问题以及置信区间和置信带的构造等问题.

Bootstrap 方法最初是由 Efron(1979)[5][6] 发明的, 随后进行了两次重要的发展. Efron(1982)[7] 的专论拓展了 1979 年的议题, 并提出了一些新的思想. 统计观点上的 Bootstrap 的说明可以参见 Efron&Gong (1983)[8] 、Efron&Tibshirani (1986)[9] 以及 Hinkley(1988)[10]. Beran&Ducharme(1991)[11] 的讲义以及 Hall 的专论中给出了关于 Bootstrap 数学上的随机处理. 在 Efron (1983)[12] 、Lunneborg (1985)[13] 、 Rasmussen (1987)[14] 以及 Efron&Tibshirani(1991)[15] 的文献中可以查到 Bootstrap 的非技术性的描述.

当 Efron 的 1979 年的论文正式提出了 Bootstrap 时, 类似的思想也出现在不同的文章中, 其中包括 Barnard(1963)[16] 和 Hope(1968)[17] 提出的蒙特卡罗假设检验方法. 尤为值得一提的是 Hartigan(1969[18],1971[19],1975[20]) 建立的置信区间的理论.

§2.2.2 Bootstrap 方法的基本思想

Bootstrap 是一种以计算机为基础的统计推断方法, 它可以衡量统计估计的准确性, 其最大优势是可以不依赖公式直接给出方差、标准差、偏差、覆盖概率和置信区间等统计问题的结果. 利用 Bootstrap 估计标准差是 Bootstrap 应用中最重要的方面之一. 下面就以 Bootstrap 标准差为例, 介绍 Bootstrap 的基本思想.

图 2.1 是 Bootstrap 过程的一个简图. 假设观察到独立的样本数据集 $x=(x_1,x_2,\cdots,x_n)$, 待估计的统计量为样本的一个函数, 记为 $s(x)$, 我们感兴趣的是 $s(x)$ 的标准差的估计. Bootstrap 算法开始先产生大量独立的 Bootstrap 样本 $x^{*1},x^{*2},\ldots,x^{*B}$, 这 B 个 Bootstrap 样本是从原始数据集中做有放回的抽样产生得到的, 其中每个样本又有 n 个元素. 为了得到标准差估计, Bootstrap 样本的个数 B 通常取 50 到 200. 相应于每个 Bootstrap 样本都可以得到 $s(x)$ 的一个 Bootstrap 复制, 记为 $s(x^{*b})$. 例如, $s(x)$ 为样本中位数, 则 $s(x^*)$ 即为

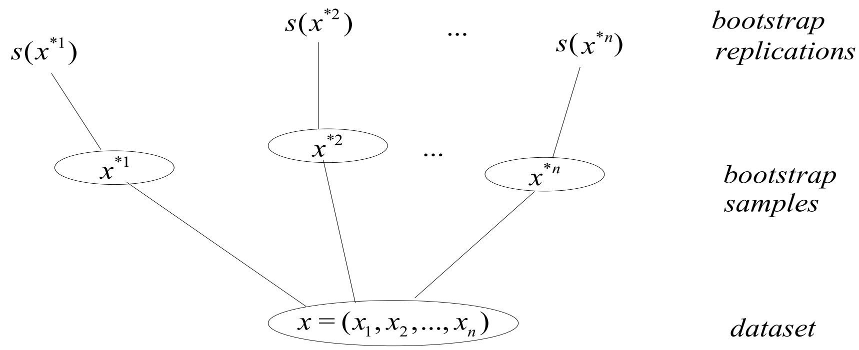

图 2.1: Bootstrap 示意图

Bootstrap 样本的中位数. 标准差的 Bootstrap 估计是 Bootstrap 复制的标准偏差，即

$$\widehat{se}_{boot} = \frac{\sum_{b=1}^{B}\left[s\left(x^{*b}\right)-s(\cdot)\right]^2}{(B-1)^{\frac{1}{2}}}$$

其中 $s(\cdot)=\frac{\sum_{b=1}^{B}s\left(x^{*b}\right)}{B}$.

由上述过程易见， Bootstrap 的基本思想包括两个步骤：首先得到样本 x 的分布 Q 的估计量 $\tilde{Q}$ ，再利用替代原则将其估计值代入待估的统计量 $s(\cdot)$ 以得到 $s(Q)$ 的估计量 $s\left(\tilde{Q}\right)$. Bootstrap 在 1979 年借助计算机运用于标准差的估计，其优点在于能够自动地完成计算，而不需要理论上的计算，且不管统计量 $s(x)$ 的数学形式有多复杂，它都可以运用.

§2.2.3 Bootstrap 置信区间

Bootstrap 标准差常用来近似参数 θ 的置信区间. 利用 Bootstrap 构造参数的置信区间的方法很多. 常见的有：

1. 标准正态置信区间

   θ 的 $1-2\alpha$ 的标准正态置信区间为 $\left[\widehat{\theta}-z^{1-\alpha}\cdot\widehat{se},\ \widehat{\theta}-z^{\alpha}\cdot\widehat{se}\right]$, 其中 $\widehat{\theta}$ 为参数 θ 的估计， $\widehat{se}$ 为标准差估计 (可以用上一节讨论的 Bootstrap 标准

差代替)， z^{α} 为标准正态分布 N(0,1) 的第 100α 个百分位点.

2. t- 区间

θ 的 $1-2\alpha$ 的 t- 区间为 $\left[\widehat{\theta}-t_{n-1}^{1-\alpha}\cdot\widehat{se},\ \widehat{\theta}-t_{n-1}^{\alpha}\cdot\widehat{se}\right]$, 其中 $\widehat{\theta}$ 和 $\widehat{se}$ 定义如上， t_{n-1}^{α} 表示自由度为 $n-1$ 的 t 分布的第 100α 个百分位点.

3. Bootstrap-t 区间

θ 的 $1-2\alpha$ 的 Bootstrap-t 区间为 $\left[\widehat{\theta}-\widehat{t}^{1-\alpha}\cdot\widehat{se},\ \widehat{\theta}-\widehat{t}^{\alpha}\cdot\widehat{se}\right]$, 其中 $\widehat{\theta}$ 和 $\widehat{se}$ 定义如上, $\widehat{t}^{\alpha}$ 定义如下: 根据上一节中 Bootstrap 方法的基本思想, 产生 B 个 Bootstrap 样本 $x^{*1},x^{*2},\ldots,x^{*B}$. 对于每个样本，计算 $z^{*}(b)=\frac{\widehat{\theta}^{*}(b)-\widehat{\theta}}{\widehat{se}^{*}(b)}$，其中 $\widehat{\theta}^{*}(b)$ 是用 Bootstrap 样本 x^{*b} 计算得到的 $\widehat{\theta}$ 的估计， $\widehat{se}^{*}(b)$ 是 $\widehat{\theta}^{*}$ 的标准方差的估计. $z^{*}(b)$ 的第 100α 个百分位点是通过 $\widehat{t}^{\alpha}$ 的值来估计的，如下式所示：

$$\frac{\sharp\{z^{*}(b)\leq\widehat{t}^{\alpha}\}}{B}=\alpha$$

例如：当 $B=1000$ 且 $\alpha=0.05$ 时， $\widehat{t}^{\alpha}$ 为 $z^{*}(b)$ 从小到大排列后的第 $B\cdot\alpha=50$ 个点.

本文采用另一种 Bootstrap 置信区间：基于统计量的 Bootstrap 分位数置信区间. 假设由原始数据集 x 产生的 Bootstrap 数据集为 x^{*}，计算得到 Bootstrap 复制 $\widehat{\theta}^{*}=s(x^{*})$. 令 $\widehat{G}$ 为 $\widehat{\theta}^{*}$ 的累积分布函数，则 $1-2\alpha$ 分位数区间定义由 $\widehat{G}$ 的 α 和 $1-\alpha$ 分位数构成：

$$\left[\widehat{\theta}_{\%,lo},\widehat{\theta}_{\%,up}\right]=\left[\widehat{G}^{-1}(\alpha),\widehat{G}^{-1}(1-\alpha)\right] \tag{2.1}$$

由于定义 Bootstrap 分布的第 100α 个百分位点为 $\widehat{G}^{-1}(\alpha)=\widehat{\theta}^{*(\alpha)}$ ，故分位数区间也可写为

$$\left[\widehat{\theta}_{\%,lo},\widehat{\theta}_{\%,up}\right]=\left[\widehat{\theta}^{*(\alpha)},\widehat{\theta}^{*(1-\alpha)}\right] \tag{2.2}$$

表达式 (2.1) 和 (2.2) 是理想 Bootstrap 的情形, 即 Bootstrap 复制数量无穷的情形. 在实际中, 只能产生有限个 (B 个) 复制. 假设产生了 B 个独立的 Bootstrap